Visual Speech Feature Extraction From Natural Speech for Multi-modal ASR

Sabri Gurbuz and John N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
E-mail:{sabrig,jgowdy}@ces.clemson.edu

Abstract

Improving the accuracy of speech recognition technology by addition of visual information is the key approach to multi-modal ASR research. In this work, we address two important issues, which are lip tracking and the visual speech feature extraction algorithm. In order to utilize the multi-modal ASR for natural speech, the visual front end algorithm must extract affine and lighting condition invariant visual speech features.

This paper focuses on both the lip tracking algorithm using the Bayesian framework and a novel pixel based visual speech feature extraction algorithm based on kurtosis measures of the frequency profile of the local image blocks. We compare the results of the proposed features with the results of outer lip contour based affine-invariant visual features, and global 2D DCT features. Experimental results in this paper are presented for a visual-only connected digit recognition task for performance comparison of the visual features.

Keywords: Lip tracking, Visual feature extraction, Kurtosis measure.

1. Introduction

The addition of visual information to audio features improves speech understanding and offers key advantages in human-computer interfaces especially in difficult environments [1–6]. Improving the existing state-of-the-art automatic speech recognition (ASR) performance by integrating the visual information of the speaker's mouth region is receiving significant attention from the speech recognition communities.

Some of the initial difficulties difficulty associated with computer lipreading (visual speech recognition) are the accurate and consistent visual region of interest (ROI) extraction, and lip tracking algorithm on the fly, which needs to be robust to a speaker's ethnic and gender variability, and other visual appearances such as glasses, facial hair, various skin color, lip color, and different lip shapes. Another difficulty difficulty is the robust and consistent visual speech feature extraction.

The development of a successful audio-visual speech recognition technology capable of adapting itself to changing environments will support both industrial and military applications. Audio-visual speech recognition research is a relatively new and advancing research area. A noise robust audio-visual speech recognition system will facilitate use of computers, increase reliability and worker productivity, and naturalize communications between human and computers. In addition, audio-visual speech recognition technology can facilitate new commercial applications such as text-driven audio-visual talking head, audio-visual speech-to-speech translation, and speech-to-video conversion for the hearing impaired.

In our earlier research [1,7], we have implemented both late integration and early (multi-stream state synchronous) integration schemes for a controlled audio-visual data set. For both integration schemes, the experimental results showed that addition of visual information improves the recognition performance. In this paper, the following objectives will be sought:

1. **Development of a lip tracking algorithm, and**
2. **A novel visual speech feature extraction algorithm that satisfies the following three criteria:**
   - i. **Affine (rotation, scale, and shear) invariance,**
   - ii. **Chrominance space shift invariance, and**
   - iii. **Chrominance space scale invariance.**

In our proposed visual speech feature extraction method, the criteria in step (i) is satisfied by affine correction, the criteria in step (ii) is satisfied by removing of the DC component of the 2D DCT coefficients, and the criteria in step (iii) is satisfied by the normalized higher order moments of the DCT coefficients of the lip image blocks.

This work is organized as follows. In section 2, we present a Bayesian framework for lip tracking, parametric formulation of the Gaussian parameters and adaptation of the parameters on the fly. Section 3 discusses the removal of affine (rotation, scale, shear) effects from the segmented lip image. In section 4, we discuss contour based affine invarinat features, pixel based normalized 2D DCT features, and describe a novel visual speech feature extraction algorithm based on kurtosis measures of the frequency profile of the local image blocks of the mouth. We present the experimental setup and the results in Section 5. Section 6 gives the concluding remarks and the proposed future work.

2. Lip Tracking Using the Bayesian Framework

The basis of the audio-visual speech recognition system is an efficient lip tracking algorithm. Computational time constraints required by applications such as audio-visual speech recognition, animated talking head design, etc., contribute to the difficulty of the task. Most lip tracking algorithms build upon the eigenspace based face detector and an ensemble of feature detectors which are used to extract pre-specified landmarks such as nostrils and lip corners to

locate the ROI (mouth region) [8,9]. The deformable template and snake based methods [10,11] have also been used for this task. All techniques have reported good results, but accuracy has decreased when there are occlusion (profile view), lighting condition change, texture changes, and quick motion. The technique we propose uses color images with Bayesian framework for classification which requires the estimation of the *a priori* probabilities and class conditional density models. The class conditional density and *a priori* probability estimation processes are described in the following sections.

In the lip tracking problem there are two distinct classes, *lip* and *non-lip*. Therefore, in this section, the two class classification problem is discussed because each sample in the image frame either belongs to *lip* class, w_1 or *non-lip* class, w_2. The conditional density functions and the *a priori* probabilities are estimated using the training data that may require extensive search to locate the *lip* and *non-lip* regions in the first frame in practice which will not be discussed here. The Bayes decision rule determines whether an observation, x, belongs to w_1 or w_2. One of the most commonly utilized probability density functions in practice is the Gaussian density function due to its computational simplicity and because it models a large number of cases in nature. The Gaussian parameters are estimated parametrically using the information from the previous frame on the fly which leads to an adaptive real time lip tracking and segmentation algorithm.

2.1. Parametric Formulation of Gaussian Density from Sample Data

In the parametric formulation of the multivariate Gaussian density, estimation of the mean vector and covariance matrices of the two classes, w_1 and w_2, are required. Let N be the number of samples drawn from a class, w_i, with respect to x in the *n-dimensional* feature space. Then the general multivariate Gaussian (normal) density given by

$$p(\mathbf{x}|w_i) = \frac{1}{\sqrt{(2\pi)^n ||\Sigma_i||}} exp\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\}, \quad (1)$$
$$i = w_1, w_2.$$

where $\mu_i = E[\mathbf{x}]$ is the mean value of the class w_i, and Σ_i is the $n \times n$ covariance matrices defined as

$$\Sigma_i = E[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T] \quad (2)$$

$||\Sigma_i||$ represents the determinant of Σ_i and $E[.]$ is the expected value of a random variable. The parameters μ_i and Σ_i can be estimated without bias by the sample mean and sample covariance matrix as

$$\hat{\mu_i} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j^{(i)}, \quad i = w_1, w_2 \quad (3)$$

$$\hat{\Sigma_i} = \frac{1}{N-1}\sum_{j=1}^{N} (\mathbf{x}_j^{(i)} - \mu_i)(\mathbf{x}_j^{(i)} - \mu_i)^T, \quad i = w_1, w_2 \quad (4)$$

where $\mathbf{x}_j^{(i)}$ is the jth sample vector from the ith class.

2.1.1. Class Conditional Mixture Density Estimation

Given the data sets for *lip* and *non-lip* classes from the previous frame, we can form the class conditonal mixture density function in general as follows.

1. Form a 6-dimensional attribute data set for each class from color and texture measures (R, G, B, R_v, G_v, B_v) for each pixel location, and cluster it (possibly into three clusters for lip, tongue, and teeth) using an unsupervised K-means clustering algorithm.
2. Form the parametric class conditional density models $P(x \mid w_L^{(i)})$ using the method described in Section 2.1 for each cluster, where i represents the cluster i.d.
3. Similarly, repeat step 2-6 to form the parametric class conditional density models $P(x \mid W_{nL}^{(i)})$ for non-lips (nL).
4. Form the conditional density mixture models using weighted sum of the conditional densities belonging to clusters. That is,

$$P(\mathbf{x} \mid w_i) = \sum_{m=1}^{C} c_m P(\mathbf{x} \mid w_i^{(m)}), \quad i = L, nL \quad (5)$$

where C is the number of cluster for the lip or non-lip class, and $c_m = n_m/N$ is the mixture weight obtained by taking the ratio of the number of pixels in cluster m to total number of pixels in that class.

2.1.2. A Priori Probability Estimation

As shown in Equation 10, *a priori* probability specification is an important task for a Bayesian classifier since the *threshold value* of the likelihood ratio is based on the *a priori* class probabilities. Basically, it is desired to obtain a speaker and time (frame) dependent Bayesian parameter set to adapt the skin tone color variations and lighting variations on the fly. The selection of the sample data for obtaining class mean vectors and covariance matrixes has direct effect on the parametric representation of the class conditional density models. Calculating the *a priori* class probabilities based on the number of pixels in each class data is biased to the sample data so it would be a poor choice. By careful examination of the multi-variate Gaussian density function in Equation 1, one intuitional choice of the *a priori* class probabilities would be biasing them to determinant of the covariance matrixes of the classes, as

$$p(w_i) = \frac{||\Sigma_i||}{||\Sigma_1|| + ||\Sigma_2||}, \quad i = w_1, w_2 \quad (6)$$

where $p(w_1) + p(w_2) = 1$. Figure 1 shows the class regions based on the *threshold value* of the likelihood ratio (Bayes decision rule) and the effect of *a priori* class probability selection.

2.2. Bayesian Decision Rule

Let x be an observation vector (a set of features belong to a pixel location in the image frame). Our goal is to design a Bayes classifier to determine whether x belongs to w_1 or w_2. The Bayes test using *a posteriori* probabilities may be written as follows:

$$p(w_1 \mid \mathbf{x}) \underset{w_1}{\overset{w_2}{\lessgtr}} p(w_1 \mid \mathbf{x}), \quad (7)$$

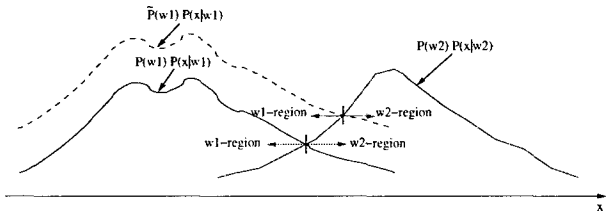


Figure 1: Bayes decision rule and the effect of the a *priori* class probability values.

where $p(w_i \mid \mathbf{x})$ is *a posteriori* probability of w_i given x. Equation 7 shows that, if the probability of w_1 given x is larger than the probability of w_2, then x is declared belonging to w_1, and vice versa. Since direct calculation of $p(w_i \mid \mathbf{x})$ is not practical, we can re-write the *a posteriori* probability of w_i using the Bayes theorem in terms of *a priori* probability and the conditional density function $p(\mathbf{x} \mid w_i)$, as

$$p(w_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid w_i)p(w_i)}{p(\mathbf{x})} \tag{8}$$

where $p(\mathbf{x}$ is the mixture density function, and is positive and constant for all classes. Then, the decision rule shown in Equation 7 can be written as

$$p(\mathbf{x} \mid w_1)p(w_1) \underset{w_1}{\overset{w_2}{\lessgtr}} p(\mathbf{x} \mid w_2)p(w_2) \tag{9}$$

or re-arranging both sides, we get

$$L(\mathbf{x}) = \frac{p(\mathbf{x} \mid w_1)}{p(\mathbf{x} \mid w_2)} \underset{w_1}{\overset{w_2}{\lessgtr}} \frac{p(w_2)}{p(w_1)} \tag{10}$$

where $L(\mathbf{x})$ is called the *likelihood ratio*, and $p(w_2)/p(w_1)$ is called the *threshold value* of the likelihood ratio for the decision. As shown in Equation 10 *a priori* probability specification is an important task for a Bayesian classifier. Because of the exponential form of the involved densities in Equation 10, it is preferable to work with the monotonic functions called discriminant functions following discriminant functions obtained by taking the logarithm of both sides of the Equation shown in 9.

$$q_i(\mathbf{x}) = ln(p(\mathbf{x} \mid w_i)p(w_i)), \text{ or} \tag{11}$$

$$q_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i) + \ln p(w_i) + c_i \tag{12}$$

where $c_i = -(1/2)\ln 2\pi - (1/2)\|\Sigma_i\|$ is a constant. In general Equation 12 has a nonlinear quadratic form and using Equation 12, the Bayes rule is as follows, which is preferable for the efficiency of calculation speed.

$$q_1(\mathbf{x}) \underset{w_1}{\overset{w_2}{\lessgtr}} q_2(\mathbf{x}). \tag{13}$$

2.3. Lip Tracking Algorithm and ROI Selection

The Bayesian framework descibed in this paper utilizes color images with no prior labeling. The goal is to segment the lip region in the current frame and select the ROI for the following frame to limit the search space. The basic lip tracking and ROI selection procedures are described below.

- Obtain $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$ using Equation 11 for every pixel in the image.
- Use an averaging filter on the $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$ to obtain $\{S_1(x)\}$ and $\{S_2(x)\}$. The smoothing operation reduces the noise effect.
- Apply the Bayesian classification rule to every pixel in the image frame to obtain binary lip candidate pixels, as

$$S_1(\mathbf{x}) \underset{w_1}{\overset{w_2}{\lessgtr}} S_2(\mathbf{x}). \tag{14}$$

- Segment the lip region (using the heuristics such as largest region between nostrils and chin) in the binary image resulted from the Bayes classifier.

The Bayesian classifier is applied to the full image array for the first frame. But once the lip region is detected on the current frame, the next frame's search space is bounded by a rectangular ROI, obtained by enlarging the current lip region by 25% of width and height in vertical and horizontal directions, respectively. Thus, the Bayesian classifier is applied to the ROI on the next frame to enable the real time lip tracking instead of the full image array search.

Adapting classifier parameters on the fly makes algorithm more robust to lighting changes between frames. Also the initial color information extracted from the first image frame may have several problems with changing conditions. Firstly, the color features obtained for a person by a camera is influenced by the ambient lighting conditions and orientation of the speaker's face during speech. Secondly, different cameras produce significantly different color features even for the same person under same lighting conditions. Our work aims to overcome this difficulty by adapting the classifier parameters on the fly using the information from the previous frame. The procedure is described as

- Extract the color features for *lip* class.
- Extract the color features for *non-lip* class.
- Update the classifier parameters using the data obtained from above two steps.

3. Removing Affine Parameters from Lip image

In the audio-visual speech and speaker recognition task, both contour based and pixel based visual features need to be independent from the affine (rotation, scale, shear and translation) parameters. In order to utilize the audio-visual speech and speaker recognizer for natural speech, the lip image for every frame needs to be pre-processed for removing the affine parameters before the visual feature extraction process described in the following sections is applied. Then, a question can be posed whether if affine (rotation, scale, shear and tranlation) parameters convey linguistic information to utilize for the recognition task.

3.1. Lip-Rotation Problem

Lip-rotation correction on the fly for natural speaker movement is essential for robust audio-visual speech and speaker recognition. Utilizing lip corners or some other facial features such as nostrils and eye corners may be problematic for rotation correction due to the complexity of locating such facial features accurately during natural speech [9, 12].

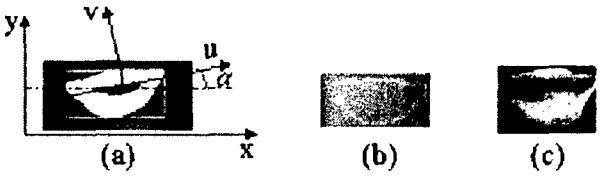


Figure 2: Lip rotation correction: a) rotation correction using the PCA, b) outer lip contour after rotation correction, c) gray lip image after ration correction and scaling to 96x64 pixels.

We propose a principal component analysis (PCA) based rotation estimation and correction method to overcome the difficulties mentioned above. Jump

3.1.1. Rotation Correction Using PCA

Principal component analysis (PCA) is a method for analyzing multivariate data to identify a set of new orthogonal axes known as principal components. The first principal component is the axis that describes most variance of the data, the second principal component is the orthogonal axis that describes the second most variance of the data, and so on. PCA is also called the Hotelling transform or Karhunen-Loéve expansion [13].

Let $\mathbf{x} = [x_1 x_2]^T$ be a 2-dimensional random variable with mean m_x and covariance matrix C based on N samples of a lip image pixel locations. The mathematical representation of PCA as follows.

$$m_{xk} = \frac{1}{N}\sum_{i=1}^{N} x_{ki}, \quad k = 1, 2 \text{ so} \tag{15}$$

$$m_x = [m_{x1}\, m_{x2}]^T \quad \text{and} \tag{16}$$

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)(x_i - m_x)^T, \tag{17}$$

where T represents the transpose operation. The task is to find the new set of orthogonal axes and estimate the rotation angle with the standard coordinate system, and then undo the rotation of the lip pixel coordinate data. Figure 2 shows the rotation correction using the PCA coordinate rotation.

In order to estimate the rotation angle α between x-axis and u-axis shown in Figure 2a, we solve for the eigenvalues $\{\lambda_1, \lambda_2\}$ of the covariance matrix C and find the eigenvector e_1 corresponding to the largest eigenvalue. The process is as follows:

$$|C - \lambda I| = 0, \tag{18}$$

and then find the eigenvectors (also called proper vector or characteristic vector), calculated as

$$C\, e_i = \lambda_i e_i, \quad i = 1, 2 \tag{19}$$

where $e_1 = [e_{x1}\, e_{y1}]^T$. The eigenvector belongs to largest eigenvalue defines the rotation angle α, as

$$\alpha = atan(e_{y1}/e_{x1}). \tag{20}$$

Then the rotation correction matrix R^{-1} can be written as

$$R^{-1} = \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix}. \tag{21}$$

Figure 3: An example of the scaling problem due to speaker's distance to camera or speaker's lip physical dimensions.

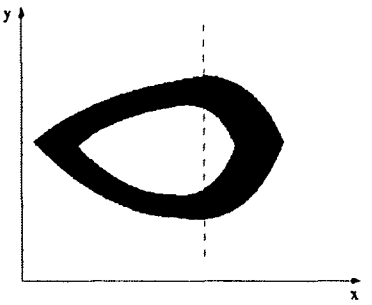


Figure 4: An illustration of the shearing in the horizontal direction.

The rotation corrected lip image is obtained by multiplying R^{-1} with the coordinates of lip pixel locations, as

$$\begin{bmatrix} x'_n \\ y'_n \end{bmatrix} = R^{-1} \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \quad n = 1, 2, \ldots, N \tag{22}$$

where $(x_n, y_n)^T$ represents the cartesian coordinates of the lip pixel locations, and $(x'_n, y'_n)^T$ represents the cartesian coordinate of the lip pixel locations after the rotation correction. Figure 2c shows the orientation of the lip shape after rotation correction and scaling of lip shown in Figure 2a.

3.2. Scaling Problem

The scaling problem occurs due to the speaker's distance to camera, the camera zoom factor and the speaker's actual lip dimensions. In this case, any pixel based visual feature extraction method such as DCT or wavelet transform method which utilizes the frequency content of the lip image may generate inconsistent (noisy) observation vectors. To overcome this problem, we propose to interpolate every lip image to same size, $N \times M$. Figure 3 shows the scaling problem example for two different speakers and the lip images of them after interpolation (scale correction).

3.3. Shearing (Uneven Scaling) Problem

Shearing occurs when the speaker's head position is not perpendicular to camera optical axis. For example, one side of the lips which may look larger than the other. Solving the shearing problem using the single 2D image information is not theorotically possible. There can be various practical approaches to minimize the shearing effect such as using the symmetry information of the lips may enable us to estimate the shear matrix by utilizing the least squares estimate method and undo the shearing. Figure 4 illustrates a typical example of a shearing effect in the horizontal direction.

The shearing may also be associated with the accent of a speaker, depending on certain visimes. Then, the similar

question can be posed whether shearing conveys a linguistic information.

4. Visual Speech Feature Extraction

Lipreading clearly meets at least two practicable criteria: It mimics human visual perception of speech recognition, and it contains information that is not always present in the acoustic signal [3,4,14–16]. Petajan is one of the first researchers who built a lipreading system using oral-cavity features to improve the performance of an acoustic ASR system [17]. Silsbee et al. [18] utilized vector quantization (VQ) of acoustic and visual data for their HMM based audio and video subsystems. Teissier et al. [19] utilized 20 FFT based 1-bark wide channels between 0 and 5 Khz for acoustic features and inner lip horizontal width, inner lip vertical height and inner lip area for the visual features. Chiou et al. [20] utilized active contour modeling to extract visual features of geometric space, the Karhunen-Loève transform (KLT) to extract principal components in the color eigenspace, and HMMs to recognize the combined video only feature sequences. Potamianos et al. [14,21] used Fourier descriptor magnitudes for a number of Fourier coefficients, width, height, area, central moments, normalized moments as contour features, image transform features, and hierarchical discriminant features.

In order to utilize audio-visual ASR for natural speech in varying lighting conditions, the visual front end algorithm that extracts the visual features must satisfy the three criteria presented in Section 1. The contour based feature described in Section 4.1 satisfy step (i) in the Fourier domain and is relatively independent of step (ii) and step (iii). For pixel based visual feature extraction methods, step (i) is explained in Section 3. Steps (ii) and (iii) are explained for both 2D DCT based visual features and kurtosis measure based visual features which are described in Sections 4.2, and 4.3, respectively.

4.1. AI-FDs Based Visual Features

In general, for the video feature extraction, the relationship between observed parametric outer-lip contour data x and parametric reference data $\mathbf{x}^{\circ}$ can be written as,

$$\mathbf{x}[\mathbf{n}] = A\mathbf{x}^{\circ}[\mathbf{n}+\tau] + \mathbf{b}, \tag{23}$$

where A represents a 2×2 arbitrary affine matrix, $det(A) \neq 0$, that may have scaling, rotation, and shearing affect, b represents a 2×1 arbitrary translation vector, and τ is starting point. These are removed in the Fourier domain [7,22]

The video feature extraction algorithm extracts twelve affine-invariant Fourier descriptors (AI-FDs) of the parametric outer lip contour data as well as four affine-invariant oral cavity features which are width, height, ratio of width to height, and outer lip's inner area by normalizing the next frame's corresponding oral cavity features. Dynamic coefficients, which are used as a video observation features, are obtained by differencing the consecutive image sequence features.

4.2. Normalized 2D DCT Based Visual Features

The Discrete Cosine Transform is one of the many transform methods that transforms its input into a linear combination of weighted basis functions. The 2D DCT on a NxN lip image can be written as

$$Y = C^T X C \tag{24}$$

where X is an NxN lip image, Y contains the NxN DCT coefficients, and C is an NxN transform matrix defined as

$$C_{mn} = k_n \cos\left[\frac{(2m+1)n\pi}{2N}\right], \text{ where} \tag{25}$$

$$k_n = \begin{cases} \sqrt{1/N} & \text{when } n = 0, \\ \sqrt{2/N} & \text{otherwise} \end{cases}$$

and $m, n = 0, 1, ..., N-1$. Our goal is to extract visaul features satisfying step (ii) and step (iii), and most relevant information of the lip shape from the NxN DCT coefficients. Let I^o and I be lip shape images which differ in a scale and shift factors (lighting condition). i.e.,

$$I = \alpha I^o + \delta, \tag{26}$$

where α and δ are scale and shift factors in the acceptable range[1] of the chrominance/luminance space.

From Equation 25, we know that the zeroth coefficient of the DCT transform contains the DC information (δ in Equation 26) which doesn't convey any shape information. It is also known that DCT is a linear transform and the scale factor α just scales all the DCT coefficients. So normalizing all the coefficients in the DCT domain by a coefficient Y_{mn} makes the DCT transform scale independent. Then, 35 coefficients from the lower frequencies are selected excluding the DC information. Figure 5 shows the normalized 2D DCT based visual feature extraction process.

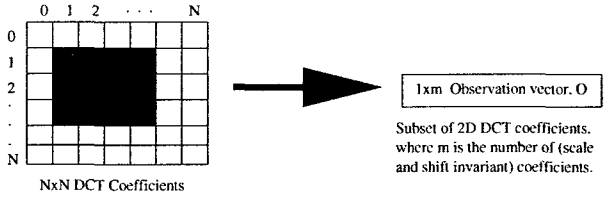


Figure 5: Normalized 2D DCT based visual feature extraction.

4.3. 2D Kurtosis Measure of the Probability Density Distribution of the DCT Coefficients

After the rotation correction and size normalization of the lip image, the resulting lip image is divided into 16 × 16 sub-blocks with 50% overlapping or non-overlapping sub-blocks, and then the two-dimensional DCT of the each block is calculated. For simplicity, let Y be the matrix of 16×16 DCT coefficients. Y(0,0) depends only on the chrominance/luminance space *shift* shown in Equation 26, and conveys no shape information. Thus, the $Y(0,0)$ coefficient is removed. The remaining coefficients are now only chrominance *space* scale dependent (see Equation 26). We remove the dependency on the chrominance space *scale* by calculating the 2D kurtosis of the frequency profile (probability distribution of DCT coefficients) of each block in the lip image discussed in the following sections. Figure 6 shows the pixel

[1] Reference and observed lip image contents are clearly visible for a range of α and δ.

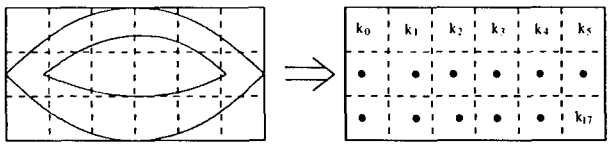


Figure 6: Illustration of FPM visual feature extraction (k_i is an appearance based visual coefficient for the ith lip image block).

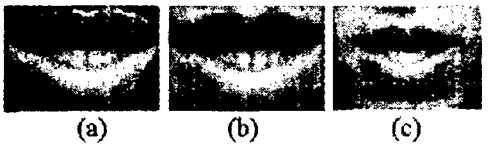


Figure 7: In search of the lip region type with 96x64 pixel size to extract visual speech features: a) exact lip region, b) exact rectangular lip region, c) extended rectangular lip region.

based visual front end process, where $k_0, k_1, \ldots, k_{17}$ are coefficients for the pixel (appearance) based visual features of the lip image. In this work, we will refer these pixel based features as frequency profile measures (FPMs), which are 2D kurtosis measures of the probability density distribution of the DCT coefficients.

In the theory of probability, the classical measure of the non-Gaussianity of a random variable is the kurtosis measure. Kurtosis measures the departure of a probability distribution from the Gaussian (normal) shape[2]. Kurtosis is dimensionless ratio, and greater than zero for most non-Gaussian random variables[3]. Specifically, for a given 2D image block function $I(n,m)$, where $m, n = 0, 1, \ldots, N$, the corresponding 2D DCT coefficients $Y(x,y)$ can be obtained as described in Section 4.2, where x and y are the spatial frequencies in the DCT domain. The high-frequency DCT coefficients[4] are discarded to minimize the video noise effect which is discussed in Section 4.3.1. The rest of the lower frequency DCT coefficients $Y(x,y)$ for $x, y = 1, 2, \ldots N/2$, are normalized to form the bi-variate probability density function $p(x,y)$. Using the notation of [23], for a given univariate random variable x with marginal probability mass function $p(x)$, mean μ_x, and existing finite moments up to the fourth moment, then, the univariate kurtosis is defined by:

$$kurt(x) = \beta_2 = \frac{m_4}{m_2^2}, \tag{27}$$

where m_2 and m_4 are the second and fourth central moments, respectively. In general, the kth central moment is defined by:

$$m_k = E[(x-\mu_x)^k] = \sum_x (x-\mu_x)^k p(x), \tag{28}$$

where marginal density function of x is

$$p(x) = \sum_y p(x,y), \tag{29}$$

where E denotes the probability expectation [24]. If x_1 and x_2 are two independent random variables, then kurtosis has the following linearity properties:

$$kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2) \quad \text{and} \tag{30}$$

$$kurt(\alpha x_1) = \alpha^4 kurt(x_1) \tag{31}$$

where is α is an arbitrary scalar. Clearly, any scale factor in Equation 27 cancels out. Let W be a p-dimensional random vector with finite moments up to the fourth, and μ and Γ be the mean vector and covariance matrix of W, respectively. Mardia [25] proposed the p-dimensional multivariate kurtosis as:

$$\beta_{2,p} = E[(W-\mu)^T \Gamma^{-1} (W-\mu)]^2, \tag{32}$$

where T denotes the transpose of a vector. Zhang [23] used 2D kurtosis of random vectors for a sharpness measure of Scanning Electron Microscopy (SEM) images. The 2D kurtosis $\beta_{2,2}$ is calculated by

$$\beta_{2,2} = [\gamma_{4,0}+\gamma_{0,4}+2\gamma_{2,2}+4\rho(\rho\gamma_{2,2}-\gamma_{1,3}-\gamma_{3,1})]/(1-\rho^2)^2, \tag{33}$$

where

$$\gamma_{k,l} = \sum_x \sum_y (x-\mu_x)^k (y-\mu_y)^l p(x,y) / [(\sum_x (x-\mu_x)^2 p(x))^{k/2} (\sum_y (x-\mu_x)^2 p(x))^{l/2}], \tag{34}$$

$$\sigma_{xy}^2 = E[(x-\mu_x)(y-\mu_y)], \quad \sigma_x^2 = E[(x-\mu_x)^2], \tag{35}$$

and

$$\rho = \sigma_{xy}^2 / (\sigma_x \sigma_y). \tag{36}$$

The 2D kurtosis measure, $\beta_{2,2}$, is dimensionless and *scale* and *shift* invariant as seen in Equation 33. In this work, the 2D kurtosis defined in Equation 33 is calculated using the probability density distribution of the DCT coefficients of the image block function $I(n,m)$. We will refer to the $\beta_{2,2}$ measure as the frequency profile measure (FPM) of an image block. The image blocks, which have zero marginal variances of x or y, are discarded for $\beta_{2,2}$ calculation, and their FPMs are arbitrarily assigned to the $\gamma_{4,0}$ value when $\sigma_x \neq 0$ and $\sigma_y = 0$, to the $\gamma_{0,4}$ value when $\sigma_y \neq 0$ and $\sigma_x = 0$, and to -1 when both $\sigma_y = 0$ and $\sigma_x = 0$.

4.3.1. Reducing the Effect of Video Noise in FPM Visual Features

It is known that the low-frequency coefficients in the DCT of the video signal contain the large details and the high-frequency coefficients contain the finer details of the image. Video noise[5] is clearly represented in the DCT coefficients and using the full spectrum of the image leads to noisy (distorted) visual features. That is why some of the high-frequncy DCT coefficients were discarded in the calculation of FPM of the image blocks described in Section 4.3. The pixel based visual front end research requires further investigation on how to minimize the effects of video noise and the dependence of FPM on the selection of the cut-off frequency.

[2]The smaller the kurtosis, the flatter the top of the distribution.

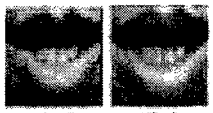
[3]Kurtosis is 3 for any univariate Gausain distribution.

[4]The upper half of the DCT coefficients are discarded.

[5]Motion blur, coding artifacts, quantization errors, electronic noise, etc., are considered to be video noises.

(a) (b) (c)

Figure 8: In search of the lip region type with 80x48 pixel size to extract visual speech features: a) exact lip region, b) exact rectangular lip region, c) extended rectangular lip region.



(a) (b)

Figure 9: Effect of interpolating on pixel based visual feature extraction: a) re-interpolated from 96x64 pixels to 60x60 pixels, b) re-interpolated from 80x48 pixels to 60x60 pixels.

Table 1: Visual-only recognition accuracy for connected digit task using the subset of the normalized 2D DCT features, FPM features, and concatenated AI-FDs and FPM features. (LR: lip region, R-LR: rectangular LR, ER-LR: extended R-LR, bl.: blocks).

Sub. of norm. 2D DCT using	TR V%	TS V%
exact LR with ini. 80x48 pixels	22.40	21.60
exact LR with ini. 96x64 pixels	23.00	20.80
R-LR with ini 80x48 pixels	24.60	17.20
R-LR with ini. 96x64 pixels	24.00	19.60
ER-LR with ini. 80x48 pixels	22.80	24.40
ER-LR with ini. 96x64 pixels	21.60	21.60
FPMs using		
exact LR with overlapping bl.	41.80	19.60
exact LR with non-overlapping bl.	35.00	24.00
R-LR with overlapping bl.	38.80	23.60
R-LR with non-overlapping bl.	34.60	22.00
ER-LR with overlapping bl.	39.00	22.00
ER-LR with non-overlapping bl.	34.20	19.60
Concat. AI-FDs and FPMs using		
only AI-FDs	18.55	21.33
exact LR with overlapping bl.	19.20	18.40
exact LR with non-overlapping bl.	17.60	18.40
R-LR with overlapping bl.	18.40	20.40
R-LR with non-overlapping bl.	17.40	18.40
ER-LR with overlapping bl.	18.40	17.60
ER-LR with non-overlapping bl.	17.80	18.80

5. Visual-Only Experimental Setup and Results

This paper discusses visual modality speech recognition (lipreading) system setup and results. The HMM states were modeled with continuous density Gaussians with single mixture components. The aim of this work is to investigate an affine and lighting conditions invariant visual feature extraction method. Therefore, the HMM model structure was kept basic. The HMM implementation was word level, left-to-right with no skip transitions with ten (eight emitting and two non-emitting) states, and diagonal covariance Gaussian mixture components since we assume that the coefficients in the observation vectors are naturally independent. All the model parameters were initialized using the Viterbi training algorithm and re-estimated using the Baum-Welch re-estimation algorithm. Viterbi recognition (dynamic programming) algorithm is utilized for the recognition.

The Clemson University Audio-visual Experimental (CUAVE) connected and continuous audio-visual digit database, which is a thirty six subject dataset, was utilized for the experiment. The visual-only experimental results are presented for a connected audio-visual digit recognition task. The following visual features from exact lip region, exact rectangular lip region, and generous rectangular lip region as shown in Figures 8 and 9 are utilized in the visual-only speech recognition system.

1. **Subset of normalized 2D DCT features**
2. **FPM features**
3. **AI-FD features**
4. **Concatenated AI-FDs and FPM features**

The subset of the 36 speaker dataset, containing 15 speakers each is uttering five times 0-9. The speakers are split into training (TR) and testing (TS) set of ten and five subjects, respectively, leading to speaker independent visual only recognition system. The results are shown in Table 1.

6. Concluding Remarks and Future Work

Table 1 shows the visual-only connected digit recognition results, where TR corresponds to training set performance and TS corresponds to test set performance, for various visual features discussed in this paper. The subset of the normalized 2D DCT features based on the training set results from exact rectangular lip region gives better results than the exact lip region and extended lip region (see in Figure 9). Another observation is that slight change in lip image content due to the linear interpolation has effects on the system's performance.

In the results obtained using FPM features, the training set performance is much better than the test set performance. Similar performance behavior was observed for a speaker dependent recognition task. Therefore, we conclude that FPM based features are highly video noise sensitive. The overlapping block based FPM features outperformed the non-overlapping block based FPM features significantly in the training set. Among the three different lip regions shown in Figure 9, the exact lip region with overlapping blocks method outperforms the results of outer two regions.

In the results obtained using concatenated AI-FDs and FPMs. the training set and test set performances are close

to each other and worse than FPMs-only results. Therefore, we conclude that each feature should be treated as a separate stream and weighted properly to bring the additional information from one another. Similarly, the slight performance increase due to the overlapping block of FPM features over non-overlapping block based FPM features can be noticable.

We also report that the number of mixtures in the Gaussian mixture model (GMM) selection and teh number of states in the silence model affects the performance of visual-only system. For example, setting GMM to twelve and using embedded training of the FPM based visual only system achieved 98% recognition accuracy on the training set, but about 16% on the speaker independent test set (which is less than the result of single GMM reported in Table 1. The similar behavior is observed for the speaker dependent set. That is, the system is being well trained with the FPM features, but the both test sets are behaving like an unmatched system due to the resulting noisy observations.

We conclude that visual noise is an important factor in visual speech feature extraction, and overlapping local image block based FPM features outperform normalized 2D DCT features, AI-FD features, and concatenated AI-FDs and FPM features. Future work will include initial lip segmentation for the Bayesian framework training and further study on the noise robust FPM feature extraction.

7. References

[1] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," in *Proceedings of ICASSP*, 2002.

[2] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speech Reading by Humans and Machines, D. G. Stork and M. E. Hennecke Eds. Springer, Berlin*, 1996.

[3] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. V. Lee, and D. Terzopoulos, "The dynamics of audio-visual behavior in speech," in *Speechreading by Man and Machine: Data, Models and Systems, D. G. Stork and M. E. Hennecke Eds. NATOSpringer-Verlag, New York, NY (1996)*, 1996, vol. 150.

[4] S. Nakamura, "Fusion of audio-visual information for integrated speech recognition," in *Audio- and Video-Based Biometric Person Authentification*, 2001.

[5] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A., Mashari, and J. Zhou, "Audio-visual speech recognition," in *JHU Workshop 2000. http://www.clsp.jhu.edu/ws2000/*, 2000.

[6] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, 1998.

[7] S. Gurbuz, , E. Patterson, Z. Tufekci, and J. Gowdy, "Lip-reading from parametric lip contours for audio-visual speech recognition," in *Proceedings of Euro Speech*, 2001.

[8] A. W. Senior, "Face and feature finding for face recognition system," in *Proceedings of AVBPA*, 1999, pp. 154–159.

[9] G. Iyengar, G. Potamianos, C. Neti, T. Faruquie, and A. Verma, "Robust detection of visual roi for automatic speechreading," in *IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 79–84.

[10] A. Yuille, "Feature extraction from faces using deformable templates," *Int. Journal of Computer Vision, 8(2)*, pp. 99–111, 1992.

[11] M. Kass, A. Witkin, and D. Terzoplpoulos, "Snakes: active contour models," in *Int. Proc. 1st Int. Conf. on Computer Vision*, 1987, pp. 259–268.

[12] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing (accepted for publication)*, 2002.

[13] A. K. Jain, *Fundamentals of Digital Image Processing*, Printice-Hall, 1997.

[14] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm-based automatic lipreading," in *Proceedings of ICIP*, 1998.

[15] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proceedings of ICASSP*, 1993.

[16] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based asr," in *Speechreading by Humans and Machines: models, systems, and applications, NATO ASI Series. Series F, Computer and Systems Sciences no. 150, pp. 461-471*, 1996.

[17] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *In CHI 88, pp. 19-25*, 1988.

[18] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech, and Audio Processing*, vol. 4, no. 5, 1996.

[19] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 6, 1999.

[20] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, 1997.

[21] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *Proceedings of ICASSP*, 2001.

[22] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition," in *Proceedings of ICASSP*, 2001, vol. 1, pp. 177–180.

[23] N. F. Zhang, M. T. Postek, R. D. Larrabee, A. E. Vladar, W. J. Keery, and S. N. Jones, "Image sharpness measurement in scanning electron microscope - part iii," in *Scanning*, 1999, vol. 21, pp. 246–252.

[24] A. Popoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1991.

[25] K.V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, pp. 519–530, 1970.